

ARTICLE



Does comprehension of L2 television programs improve through regular classroom viewing?

Geòrgia Pujadas*, *University of Western Ontario, University of Barcelona*
Stuart Webb, *University of Western Ontario*

Abstract

This study investigated second language (L2) comprehension across nine episodes of the same TV series and explored whether comprehension improves through regular classroom viewing. 121 intermediate and advanced EFL learners viewed the series under two conditions: captioned and uncaptioned. Significant differences in comprehension across episodes for both conditions were found. While there was an increase in comprehension from the first to the last episode, the improvement was not linear, and large variability in comprehension scores indicated that comprehension may be contingent to individual episodes. This suggests that the findings of studies investigating the comprehension of one episode of a TV program may not reflect the comprehensibility of related and unrelated programs. Moreover, it suggests that comprehensibility of one episode of a TV program, may not occur to the same degree in the next episode. Results also confirm earlier findings indicating significantly better comprehension when viewing with captions than without captions, and that learners with greater vocabulary knowledge achieve higher comprehension rates, regardless of the viewing condition. Lexical coverage, used as an indicator of lexical difficulty of the episodes, was negatively correlated with comprehension scores, suggesting that other factors such as prior vocabulary knowledge may be better predictors of viewing comprehension.

Keywords: *Comprehension; Audiovisual Input; Captions; Television*

Language(s) Learned in This Study: *English*

APA Citation: Pujadas, G., & Webb, S. (2025). Does comprehension of L2 television programs improve through regular classroom viewing? *Language Learning & Technology*, 29(1), 1–24.
<https://hdl.handle.net/10125/73605>

Introduction

Research investigating the value of viewing television for second language (L2) learning has indicated that audio-visual input contributes to the improvement in knowledge of grammar (for example, Lee & Révész, 2018; Pattenmore & Muñoz, 2020), pronunciation (for example, Wisniewska & Mora, 2020), and vocabulary (for example, Pujadas & Muñoz, 2019; Rodgers & Webb, 2020). This suggests that television may be an important resource for L2 learning, both within and outside the classroom setting. However, the extent to which L2 television is likely to be viewed by foreign language (FL) learners outside of the classroom is likely to depend on the extent to which it is understood.

An inherent advantage of TV series over listening or reading activities is the presence of images, which allow learners to construct meaning through an additional source of non-verbal information that can be processed automatically and independently of the learners' L2 level. Images allow learners to infer the meaning of unknown words, especially when there is temporal proximity between words and image (Peters, 2019; Pujadas & Muñoz, 2023), and consequently it becomes easier to understand the content (Durbahn et al., 2020). The theoretical foundation behind this positive interaction lies in Mayer's (2001) cognitive theory of multimedia learning, which states that learning can be improved when visual and aural information are presented simultaneously, and when learners can organize information in their working

* **Corresponding Author:** Geòrgia Pujadas, georgia.pujadas@ub.edu

memory, build connections between both sources, and integrate the new information with their prior knowledge.

Studies investigating comprehension of audio-visual input have shown that FL learners can understand L2 television to varying degrees, with an average comprehension of up to 75% for uncaptioned videos (Lee et al., 2021), up to 85% for L2-captioned videos (Li, 2014), and up to 93% for L1-subtitled videos (Birulés-Muntané & Soto-Faraco, 2016). Researchers agree that comprehension can be significantly aided by the addition of on-screen text (for example, Birulés-Muntané & Soto-Faraco, 2016; Markham et al., 2001; Montero-Perez et al., 2013), and that learners with a higher L2 proficiency level and/or larger vocabularies have better viewing comprehension (Lavaur & Baristow, 2011; Markham & Peter, 2003; Montero-Perez et al., 2013; Montero-Perez et al., 2014; Pujadas & Muñoz, 2020; 2024; Rodgers, 2013; Wang & Pellicer-Sánchez, 2022), suggesting a need to match learners' vocabulary knowledge to the lexical demands of the audio-visual input (Durbahn et al., 2020).

Comprehension may also be supported through regular TV viewing, because viewing consecutive episodes of the same TV program allows learners to accumulate contextual knowledge (Rodgers & Webb, 2011). While this is a viewing pattern that more accurately reflects how language learners actually watch television at home, studies have generally investigated L2 learning through viewing short video fragments (for example, Montero-Perez et al., 2013) or single episodes (for example, Birulés-Muntané & Soto-Faraco, 2016). Only four studies have investigated comprehension of multiple full-length episodes of L2 television programs (Fievez, 2020; Gesa, 2019; Rodgers & Webb 2017; Pujadas & Muñoz, 2020).

The present study aims to expand on prior research by examining comprehension across nine consecutive full-length episodes of a L2 TV series, and exploring whether comprehension improves through regular viewing—with and without captions. The study also investigates the effect of learners' prior vocabulary knowledge and the episodes' lexical demands on viewing comprehension to determine the extent to which these factors support comprehension.

Literature review

How can we make L2 television comprehensible?

The main objective of incorporating TV viewing in the FL classroom should be to support learners' comprehension so that students develop the confidence to view L2 television autonomously and increase their exposure to L2 input (Webb, 2015). TV series, however, are generally not developed for viewing by language learners, who may feel overwhelmed by the uninterrupted stream of speech, the fast delivery rate of the dialogues, and unfamiliar vocabulary (Guillory, 1988).

The addition of L2 captions can support comprehension. Captions can help with word decoding and speech segmentation (Charles & Trenkic, 2015) because they allow learners to visualize the speech stream and identify word boundaries (Winke et al., 2010), and can help disambiguate unclear pronunciation (Montero-Pérez et al., 2013). Virtually all studies comparing comprehension of captioned and uncaptioned video materials have shown an advantage of L2 captions over uncaptioned conditions for content comprehension. While the majority of studies have revealed significantly better comprehension for captioned viewing (Birulés-Muntaner & Soto-Faraco, 2016; Chung, 1999; Guillory, 1998; Hayati & Mohmedi, 2011; Huang & Eskey, 1999; Latifi et al., 2011; Li, 2014; Markham et al., 2001; Montero-Perez et al., 2013; Pujadas & Muñoz, 2024), several studies have found that this advantage is not statistically significant (Hsieh, 2020; Lavaur & Bairstow, 2011; Matielo et al., 2017; Montero-Perez et al., 2014; Rodgers, 2013; Rodgers & Webb, 2017; Wang & Pellicer-Sánchez, 2022), and one study (Lee et al., 2021) has reported slightly greater (non-significant) comprehension for the uncaptioned condition (see Appendix A for more information on the aforementioned studies' design, results and reported effect sizes).

Results from studies examining comprehension of L2 audio-visual input indicate that FL learners can understand unassisted audio-visual input to varying degrees, with studies reporting a mean comprehension between 38% (Li, 2014) and 75% (Lee et al., 2021) for uncaptioned viewing. Research examining comprehension of captioned L2 audio-visual input has reported mean comprehension rates between 50% (Markham & Peter, 2003) and 87% (Pujadas & Muñoz, 2024). The benefits of captions, however, vary largely across the studies, and the differences between captioned and uncaptioned viewing in those studies oscillate from a non-significant increase in comprehension of 2.6% (Rodgers & Webb, 2017) up to a 46.8% significant increase (Li, 2014). The discrepancy in results is likely due to differences in study designs, testing formats, and the audio-visual input selected. Studies have most commonly assessed comprehension of short video fragments (for example, Lee et al., 2021; Montero-Perez et al., 2013), and less commonly of single (Birulés-Muntané & Soto-Faraco, 2016, Dizon & Thanyawatpokin, 2021; Lai et al., 2021; Li, 2014; Matiolo et al., 2017; Vulchanova et al., 2015) or multiple full-length episodes of TV programs (for example, Pujadas & Muñoz, 2020, 2024; Rodgers & Webb, 2017). Studies investigating comprehension of individual episodes of TV programs, however, do not reflect the typical viewing pattern for TV series which tends to involve watching multiple episodes in sequence (Rodgers & Webb, 2017). Thus, research examining comprehension of individual episodes may not reveal the degree to which full seasons of TV series are understood by L2 learners (Rodgers & Webb, 2011; Webb, 2011).

Does comprehension improve with regular viewing?

Watching successive episodes of the same TV program in chronological order can support comprehension (Webb & Nation, 2017). This viewing approach (that is, narrow viewing) allows viewers to gradually accumulate background knowledge and become familiar with the setting, the characters and their accents, facilitating in turn the processing of new information in each subsequent episode. Corpus-driven research also shows that episodes of the same TV series (Rodgers & Webb, 2011) – or genre (Webb, 2011) – are likely to contain repeated low-frequency vocabulary, and have a smaller vocabulary load than unrelated episodes. However, only four studies have involved participants watching multiple episodes of the same television program, assessing potential gains over time (Fievez, 2020; Gesa, 2019; Rodgers & Webb 2017; Pujadas & Muñoz, 2020).

Rodgers and Webb (2017) investigated comprehension gains through viewing 10 full-length (42-minute-long) episodes of the same TV series with and without captions. Participants ($N = 372$) viewed one episode per week and completed comprehension tests consisting of a combination of multiple-choice, true-false and sequencing items. Results showed that comprehension varied from episode to episode – ranging from 60.1% to 73% in the captioned condition ($M: 66.56\%$), and from 53.7% to 70.8% in the uncaptioned condition ($M: 63.89\%$) with the captioned group consistently outperforming the uncaptioned group. There were, however, significant gains in comprehension over time, quickly improving after the first episode in both conditions. While the captioned condition had significantly higher comprehension at the start of the study, the difference in comprehension disappeared at the end. Rodgers and Webb suggested that, by the end of the study, the uncaptioned group had progressively accumulated enough background information to support comprehension to the point that the additional support of captions for the other group did not provide any additional benefit. Rodgers and Webb also noted that comprehension varied considerably amongst participants throughout the study, with a minimum mean comprehension of 32% and a maximum of 93.4%.

Pujadas and Muñoz (2020) examined viewing comprehension by a group of adolescent, beginner EFL learners. Participants ($N = 88$) watched 24 (20-minute-long) episodes of the same TV series (one per week), either with L2 [English] captions or L1 [Spanish] subtitles. Comprehension was assessed after each episode through multiple-choice and true-false items. Results showed that—regardless of the language—comprehension scores also varied across episodes, ranging from 71.7% to 90.2% with L1 subtitles ($M: 82.6$; $SD: 1.3$), and from 50.7% to 80% with L2 captions ($M: 64.6$; $SD: 1.6$). Contrary to Rodgers and Webb's study, no clear pattern of improvement was observed from the first to the last episode, although 74% of participants reported understanding the series better at the end of the study.

Pujadas and Muñoz suggest that 24 episodes (adding up to 8.5h) might not have been sufficient to show gains in comprehension for students at this young age (13-year-olds) and proficiency level (beginner).

Gesa (2019) also investigated comprehension of several full-length TV episodes with participants in Grade 6, Grade 10, and at university. Participants viewed 24 episodes with L1-Spanish subtitles (Grade 6) or L2-English captions (Grade 10), while university participants viewed 9 episodes with L2-English captions. Comprehension was assessed through post-viewing tests made up of multiple-choice, true-false, and sequencing items. Results showed that comprehension varied significantly between episodes, ranging from 36.97% to 65.76% in Grade 6 ($N = 23$; $M: 52.84\%$), from 60.83% to 82.76% in Grade 10 ($N = 30$; $M: 69.34\%$), and from 76.32% to 88.65% at university ($N = 38$; $M: 81.47\%$). No clear pattern of improvement in comprehension could be observed from the first to the last episode in any age group.

Finally, Fievez (2020) looked at comprehension of six (50-minute-long) TV episodes with glossed L2-French captions with young adult learners (that is, 17-20 year-old). Participants ($N = 67$) were asked to view the episodes in their free time, spaced at least over 10 days ($M: 16$ days; $Min: 11$ days; $Max: 29$ days). Comprehension was assessed after each episode through a 20-item test consisting of multiple-choice and true-false questions. Results showed that there was an improvement over time, with the last episode receiving a significantly higher number of correct responses (86.4%) than the first (81.7%) and second (82.3%) episodes. In line with Pujadas and Muñoz (2020), participants reported that they felt their listening skills improved and that they could cope better with the fast speech encountered in the series. Fievez also reported significant variation in comprehension across participants, with an average minimum score at 55.83% and an average maximum score of 96.67%.

Taken together, results from prior research indicates that comprehension can vary significantly across episodes in the same study – with the largest differences in comprehension across TV episodes ranging from 4.7% (Fievez, 2020) to 30% (Pujadas & Muñoz, 2020) – as well as amongst participants. However, the extent to which comprehension improves with regular viewing remains unclear, with two studies showing significant gains over time (Fievez, 2020; Rodgers & Webb, 2017), and two showing no clear pattern of improvement (Gesa, 2019; Pujadas & Muñoz, 2020).

How important is vocabulary knowledge for comprehension of L2 television?

Studies of viewing L2 audio-visual input have shown that learners' prior vocabulary knowledge plays a role in learning L2 vocabulary (for example, Feng & Webb, 2020; Peters & Webb, 2018; Pujadas & Muñoz, 2019), and studies of comprehension of audio-visual input concur: the higher the proficiency level and/or the larger the vocabulary knowledge, the higher the likelihood that the input is understood (Markham & Peter, 2003; Montero-Perez et al., 2013; Montero-Perez et al., 2014; Pujadas & Muñoz, 2020, 2024; Rodgers, 2013; Wang & Pellicer-Sánchez, 2022). This is not surprising, because learners with a larger vocabulary will potentially know (or be familiar with) a larger proportion of the words that appear in the episode, thus allowing them to devote their attention to the words that are unknown. For example, Montero-Perez, et al. (2013, 2014), carried out a pair of studies with L2-French intermediate to high-intermediate learners, involving short video clips with different captioning types (for example, full-captioning, key-word captioning, no captioning). Results showed that vocabulary knowledge (measured through a validated self-designed multiple-choice test assessing knowledge of the 2,000-7,000-word frequency bands) had a significant effect on comprehension independently of the type (or presence) of captions. In a study examining comprehension of 24 full-length episode of a TV series, Pujadas and Muñoz (2020) found similar results. Prior vocabulary knowledge (measured by the X_Lex test—Meara & Milton, 2003) accounted for 45% of the variance in comprehension when L2-captions were present.

To ensure that L2 TV series are understood, the lexical demands of the episodes should be considered (Webb & Nation, 2017). Studies of viewing and listening comprehension have shown that when learners know at least 90% of the words, they may demonstrate sufficient understanding of spoken (Van Zeeland & Schmitt, 2013) and audio-visual (Durbahn et al., 2020) input. There are many studies investigating the lexical demands of different types of L2 input (*see* Nurmukhamedov & Webb, 2019), but only one study

has investigated its direct effects on viewing comprehension (Durbahn et al., 2020), while two have done so for listening (Bonk, 2000; Van Zeeland & Schmitt, 2013), and three for reading (Hu & Nation, 2000; Laufer, 1989; Schmitt et al., 2011). In the case of Durbahn et al.'s (2020) study, which involved viewing clips from an uncaptioned documentary, it was found that the percentage of known words in the material correlated with comprehension scores, but that the correlation was not strong ($r_s = .29$ to $.36$). While reaching 90% lexical coverage of episodes of TV series may indicate that the episodes might be comprehensible, it does not ensure that learners will understand it (Webb, 2021). A secondary aim of the present study was to examine the degree to which prior vocabulary knowledge and lexical coverage predict comprehension. This may help teachers to more effectively select audio-visual input that is comprehensible for their students.

The present study

The aim of the present study is to investigate comprehension of nine captioned and uncaptioned episodes of one English language television program over one semester by intermediate and advanced EFL learners, and to explore whether comprehension improves over time as students watch more episodes. The study also seeks to examine the extent to which learners' prior vocabulary knowledge and the lexical demands of episodes predict comprehension of L2 television. Specifically, the present study addresses the following research questions:

1. To what extent do adults understand different uncaptioned episodes of the same television program over one semester of extensive viewing?
2. To what extent do adults understand different captioned episodes of the same television program over one semester of extensive viewing?
3. To what extent do captions improve comprehension of different episodes of the L2 television program?
4. To what extent do learners' vocabulary size, episodes' lexical coverage, and time (that is, the chronological order of the episodes) predict comprehension of episodes of the L2 television program?

Methodology

Participants

The initial pool of participants were 161 second-year university students (117 female; 44 male) from a university in Spain. They were Catalan-Spanish balanced bilinguals, and had a mean proficiency level of B2 / C1 according to the Oxford Placement Test, although the level ranged from A1 to C2. Participants came from three different BA programs (English Studies, Modern Languages and Audio-visual Communication), and all participants were enrolled in a compulsory English course; thus, data was collected from the entire population available (Brysbart & Stevens, 2018).

Participants had been randomly distributed to classes by the university, and each intact class was assigned to a different viewing condition. Two classes were assigned to the captioned group (CG) ($N = 97$) and two classes to the uncaptioned group (UG) ($N = 64$). Since the study was embedded within the regular English course, all students took part in the viewing sessions. However, participants who missed more than two episodes were excluded from the final sample, leaving a total of 121 participants (CG = 75; UG = 46). For the fourth research question, participants who had not completed the tests of prior vocabulary knowledge were also excluded from the analysis, leaving a total of 111 participants (UG = 43; CG = 68).

Audio-visual input

Nine consecutive episodes from the 5th season of the series *I Love Lucy* (Oppenheimer & Arnaz, 1951) were selected as viewing material for the study. This TV program was chosen for several reasons. First,

due to its original release date and the fact that it was not currently broadcast in Spain, it was a TV series that was relatively unknown to the participants, reducing the chance that they had seen the episodes before. Second, while the series had a season-long story arc, each episode contained a full story arc itself, which allowed viewers to gather background information while being able to follow the episodes even if they missed one or two over the term.

The episodes had a mean running time of approximately 30 minutes—including opening credits—which added up to a total viewing time of 270 minutes. By the end of the study participants had been exposed to a total of 32,374 tokens across the episodes. The vocabulary from the 9 episodes was analysed using the RANGE software (Nation & Heatley, 2002). Table 1 shows the lexical coverage provided by the first three 1000-word family levels, marginal words (MW) and Spanish words (SP) (a main character was Cuban and occasionally spoke in Spanish; this 0.86% of Spanish words were considered known by participants). The analysis of the lexical profile showed that 90% coverage was reached in 7 episodes at the 2000-level (plus MW and SP).

Table 1

Cumulative Lexical Demands (in percentage) Per Episode Up to the 3k Word List Level

	Episode								
	1	2	3	4	5	6	7	8	9
1,000 word list	85.84	86.50	81.34	85.33	86.09	86.93	87.57	82.40	82.67
Marginal words (MW)	88.74	88.13	83.38	88.09	89.57	89.56	90.34	85.92	86.16
Spanish words (SP)	89.13	88.47	87.70	88.09	89.75	89.87	90.40	86.91	86.52
2,000 word list	92.51	93.40	91.51	91.12	93.68	93.37	93.52	89.96	89.39
3,000 word list	94.08	94.59	93.43	93.03	95.66	94.48	94.90	92.21	91.41

For the present study, the coverage provided by the first 1,000-word family level (plus marginal words and Spanish words) was used as the measure of the episodes' lexical demands (that is, third row in Table 1). This was done because at this level there was a larger standard deviation amongst the episodes, potentially reflecting more variation in terms of vocabulary load. Proper nouns, which constituted a considerable percentage of running words in the series¹, were excluded, because assuming that participants were familiar with them could skew the results (Klassen, 2021; Webb, 2021).

Comprehension tests

Comprehension of the episodes was assessed by post-viewing tests administered to participants after each episode. Participants were given 10 minutes to complete each comprehension test. Each comprehension test included 10 items (5 multiple-choice and 5 true-false items), for a total of 90 items for all episodes. All items were scored dichotomously (correct or incorrect). Test items were written in the participants' L1 (Catalan or Spanish).

Comprehension questions were designed around Buck's (2001) definition of the listening construct, which describes listening as "the ability to: 1) process extended samples of realistic spoken language, automatically and in real time; 2) understand the linguistic information that is unequivocally included in the text; and 3) make whatever inferences are unambiguously implicated by the context of the passage." (Buck, 2001, p. 114). Consequently, two types of items were created: textually explicit items and inferential items. To answer textually explicit items, participants had to retrieve information that was explicitly (not necessarily literally) stated in the episode. For inferential items, participants had to

integrate different pieces of information given during the episode and infer the answer. Item type was based on Pujadas and Muñoz (2020), which in turn was an adaptation of Rodgers and Webb (2017). Each comprehension test included items in two formats (that is, multiple-choice items and true-false items), because using a variety of question formats provides a more balanced assessment (Buck, 2001). All items were designed in a way that the information provided by the images of the video alone was not sufficient to answer the questions. An item discrimination index was used as an internal validation measure (Kelley, 1939), showing that the mean discrimination index was *regular* for one test (E9 [0,27]), *good* for two (E2 [0,33], E3 [0,38]), and *very good* for the other six (E1 [0,45], E4 [0,40], E5 [0,49], E6 [0,48], E7 [0,48], E8 [0,40])². An example of item type and format can be found in Appendix B.

Tests of prior vocabulary knowledge

The X_Lex test (Meara & Milton, 2003) and Y_Lex test (Meara & Miralpeix, 2006) were used as measures of prior vocabulary knowledge (see Miralpeix, 2012 for validation evidence). X_Lex and Y_Lex are computerized tests in checklist format that provide an approximation of learners' L2 receptive vocabulary knowledge of the first 5,000 word families and the 5,001 to 10,000 word families, respectively. Each test randomly presents in a context-free environment 20 items from each 1,000-level, plus 20 pseudo-words to control for guessing (that is, scores were adjusted downwards when pseudo-words were selected, and the test was considered invalid if more than six pseudo-words were checked; Miralpeix, 2012). For the analysis, the scores of both tests were added to obtain a single measure. Participants' estimated mean vocabulary knowledge was 5,121 words (Min: 1900; Max: 8,100; *SD*: 1,240). The mean for the captioned groups was 5,180 (Min: 1950; Max: 7,950; *SD*: 1,250), and for the uncaptioned groups it was 5,029 (Min: 1900; Max: 8,100; *SD*: 1,218).

Procedure

The study took place over 11 weeks. During weeks 1 and 2, participants were informed about the study, signed a consent form if they agreed to participate in the research, completed a general proficiency test (that is, Oxford Placement Test) and the X_Lex and Y_Lex tests. From weeks 3 to 11, participants viewed the 9 episodes of *I Love Lucy*—one per week—according to the experimental condition they had been assigned to: captions or no captions. Each viewing session followed the same procedure. First, participants viewed the episode, and then completed the comprehension test. Tests were not corrected in class, and no feedback was given to students until the end of the study. Participants were asked to pay attention and try to understand the content of the episodes. They were not allowed to take notes when viewing episodes, because research indicates that notetaking can have a positive impact on L2 comprehension (Hayati & Jalilifar, 2009). Viewing sessions were conducted by the course instructors, who had been trained by the first author. It is important to note that watching television as a tool for L2 learning in classrooms in Spain has become more common in recent years with research consistently indicating positive contributions to learning (for example, Gesa & Miralpeix, 2023; Muñoz & Miralpeix, 2024; Pujadas & Muñoz, 2019, 2020, 2024). Thus, the viewing conditions presented in this study have ecological validity within L2 learning classrooms in Spain.

Data analysis

To address the first and second research questions, which explored differences in comprehension across episodes, a series of Wilcoxon signed-rank tests³ were run to compare each episode's mean score against the scores obtained in the other eight episodes, both in the uncaptioned and captioned condition. To address the third research question, which focused on the additive effects of captions on comprehension, a series of Mann-Whitney U tests⁴ were run to compare comprehension scores between viewing conditions. Finally, to address the fourth research question, a Generalized Linear Mixed Model (GLMM) with binomial distribution and logit function, and with crossed random effects for participants and items, was run to assess the potential effects of participants' vocabulary knowledge, the episodes' lexical demands, and time (that is, the chronological order of the episodes) on viewing comprehension. The GLMM (based on 9,860 observations) was run with comprehension score at the item level (correct/incorrect) as the

dependent variable, and viewing condition (captioned/uncaptioned), vocabulary knowledge (continuous), episode's lexical demands (continuous), and time (1 to 9) as the independent variables, as well as all two-way interactions. Continuous variables were centred (that is, z-scored) before analyses, and non-significant factors and interactions ($p < .05$) were removed one by one to arrive at the best fitting model (see Appendix C) and the model converged. Visual inspection of the residual plots indicated that the GLMM met the assumptions of linearity, homoscedasticity, and normal distribution of residuals.

Results

Comprehension of uncaptioned episodes

Table 2 displays the mean comprehension scores for each episode for the uncaptioned viewing condition.

Table 2

Mean Comprehension Scores by Episode for the Uncaptioned Group (UG)

	<i>N</i>	Min.	Max.	Mean	<i>SD</i>
Episode 1	45	4	10	7.51	1.39
Episode 2	46	5	10	8.30	1.43
Episode 3	44	3	10	7.25	1.70
Episode 4	42	5	10	7.48	1.38
Episode 5	44	3	10	6.70	1.89
Episode 6	46	2	10	6.48	1.97
Episode 7	45	2	10	6.73	1.59
Episode 8	45	4	10	7.87	1.55
Episode 9	46	5	10	8.89	1.40

In answer to the first research question, the mean comprehension for the uncaptioned group (UG) was 73.5%, with mean scores ranging from 64.8% (episode 6) to 88.9% (episode 9), and a maximum difference in comprehension between episodes of 24%. A Wilcoxon signed-rank test showed that the difference between these two episodes was significant, with a large effect size⁵ ($Z = 5.418$, $r = .80$, $p < .001$). Table 2 shows that there was considerable variation in comprehension scores both between and within episodes. A series of Wilcoxon signed-rank tests were run to further explore differences in comprehension across the nine episodes (that is, 36 tests comparing each episode to the other eight, see supplementary file). Results showed that there was a statistically significant difference in comprehension in 25 out of the 36 comparisons ($p < .05$), with medium (4) to large (21) size effects ($r = .31$ to $.80$). Episode 9 had significantly higher mean comprehension than the other eight episodes, and episode 2 showed a similar pattern (except when compared to episode 8, where the differences were non-significant). Regarding the variation *within* each individual episode, participants' minimum scores ranged from 20% to 50%, while the maximum score was at ceiling in all episodes (that is, 100%). Variation across participants was also considerable; the participant with the lowest score in this condition had a mean comprehension of 52.2% for the 9 episodes viewed, while the one with the highest score achieved 91.1%.

Comprehension of captioned episodes

Table 3 displays the mean comprehension scores by episode for participants in the captioned condition.

Table 3

Mean Comprehension Scores by Episode for the Captioned Group (CG)

	<i>N</i>	<i>Min.</i>	<i>Max.</i>	<i>Mean</i>	<i>SD</i>
Episode 1	75	3	10	8.20	1.52
Episode 2	72	6	10	8.94	1.00
Episode 3	75	5	10	8.60	1.17
Episode 4	73	5	10	8.97	1.31
Episode 5	73	5	10	8.51	1.26
Episode 6	75	3	10	8.40	1.32
Episode 7	74	4	10	7.91	1.30
Episode 8	75	5	10	8.93	1.15
Episode 9	75	8	10	9.31	0.73

In answer to the second research question, participants in the captioned group (CG) had a mean comprehension score of 86.3%, with comprehension ranging from 79.1% (episode 7) to 93.1% (episode 9). The maximum difference in the captioned condition between the episode with the highest and lowest comprehension was smaller (14%) than in the UG, but the difference between these two episodes was also significant, with a large effect size ($Z = 6.276$, $r = .73$, $p < .001$). A series of Wilcoxon signed-rank tests were run to assess differences amongst episodes, and results revealed that the percentage of comprehension was statistically different in 25 out of the 36 comparisons ($p < .05$), with medium (15) to large (10) effect sizes ($r = .24$ to $.73$). As in the uncaptioned condition, episode 9 had a significantly higher mean comprehension than the other eight, while episode 7 had a significantly lower mean comprehension than the rest of the episodes. The minimum score in the captioned episodes ranged from 30% to 80%, while the maximum score was again at ceiling (that is, 100%). Regarding variability within participants, the lowest mean score across episodes was 60%, whereas the highest was 96.7%.

Comparing uncaptioned and captioned viewing

The third research question addressed the additive benefits of captions for viewing comprehension. The mean comprehension score for the UG was 73.5%, while the mean comprehension score in the CG was 86.3%, indicating that on average participants' comprehension was 12.8% higher when they had access to captions. Results from a Mann-Whitney U test revealed that this mean difference was significant, approaching a large effect size ($Z = 5.968$, $r = .58$, $p < .001$).

Results revealed that the CG had higher comprehension scores than the UG for all nine episodes, with differences ranging from 4.9% (episode 9) to 19.2% (episode 6). A series of Mann-Whitney U tests⁶ confirmed that CG had significantly higher comprehension scores than the UG for all episodes (with medium effect sizes) except the last one ($Z = .864$, $r = .08$, $p = .388$) (see Table 4).

Table 4

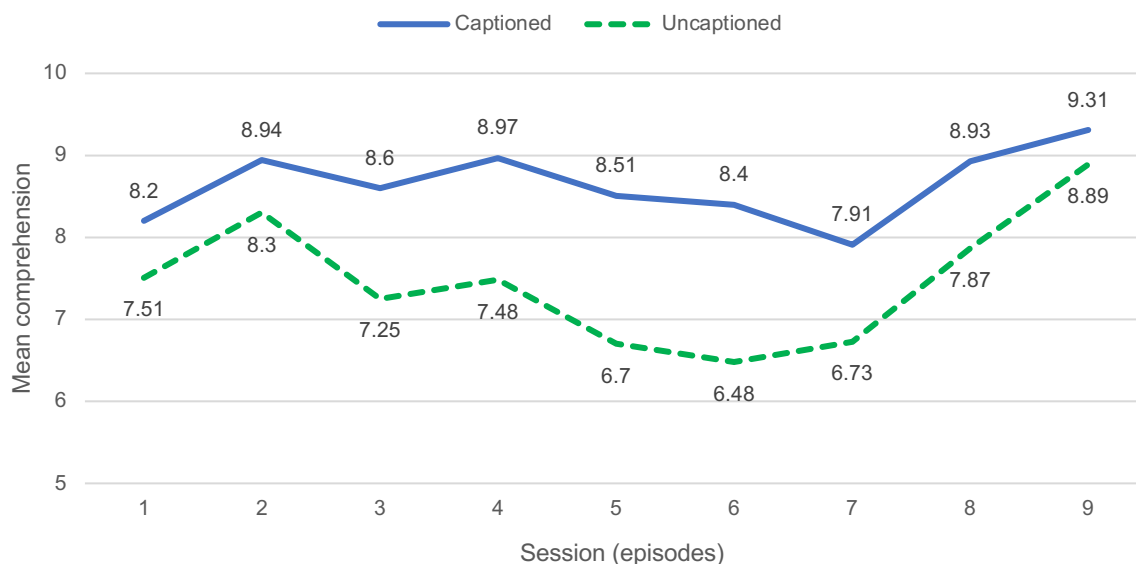
Comparison between UG and CG comprehension scores

	<i>N</i>	UG Mean (max. 10)	<i>N</i>	CG Mean (max. 10)	Contrast	<i>Z</i>	<i>r</i>	<i>p</i>
Episode 1	45	7.51	75	8.20	0.69	2.657	0.24	.008
Episode 2	46	8.30	72	8.94	0.64	2.414	0.22	.016
Episode 3	44	7.25	75	8.60	1.35	4.504	0.41	<.001
Episode 4	42	7.48	73	8.97	1.49	5.258	0.49	<.001
Episode 5	44	6.70	73	8.51	1.81	5.179	0.48	<.001
Episode 6	46	6.48	75	8.40	1.92	5.527	0.50	<.001
Episode 7	45	6.73	74	7.91	1.18	4.320	0.40	<.001
Episode 8	45	7.87	75	8.93	1.06	4.128	0.38	<.001
Episode 9	46	8.89	75	9.31	0.42	0.864	0.08	.388

While comprehension varied significantly from episode to episode, it followed a similar pattern across the two viewing conditions, particularly at the beginning and at the end of the study (see Figure 1). The episode with the highest comprehension score in both viewing conditions was episode 9, while the episode with the lowest comprehension score was episode 7 for the captioned group, and episode 6 for the uncaptioned group.

Figure 1

Mean Comprehension Scores by Episode, in the Captioned (CG) and Uncaptioned (UG) Conditions



Effects of vocabulary knowledge, lexical demands and time

Results from the GLMM indicated that having access to captions was the strongest predictor of comprehension ($p < .001$). The odds of a correct response were 2.4 times (or 41.6%) higher for participants in the captioned condition compared to the odds of participants in the uncaptioned condition, all other factors held constant. The estimated probability of selecting the correct response was 0.91 for the CG (95% CI [0.88, 0.93]) in comparison to the 0.80 in the UG (95% CI [0.76, 0.84]).

Vocabulary knowledge also emerged as a significant positive predictor of comprehension ($p < .001$), with greater vocabulary knowledge leading to higher comprehension scores. Results indicated that by increasing vocabulary knowledge by one standard deviation (*SD*), the odds of a correct response increased by 39%. The absence of interaction effects between this measure and the viewing condition implies that the effect of vocabulary knowledge was independent of captions.

Regarding the episodes' lexical demands, the model revealed that lexical coverage of the 1000-level (plus marginal words and Spanish words) had a significant negative effect on comprehension ($p < .001$), and that increasing lexical coverage by one *SD*, would decrease the probability of answering the comprehension items correctly by 0.427 log units, compared with the episodes at the grand mean of lexical coverage in this current sample. In other words, as lexical coverage increases by one *SD*, the odds of a correct response are multiplied by 0.65 (or reduced by 35%)⁷.

The variable of *time*—that is, the viewing order of the episodes, from 1 to 9—did not appear to predict comprehension in the model. An additional analysis comparing the results from the first and last episode was run to further explore the potential effects of time. Overall, episode 1 had a mean comprehension of 79.4% (UG: 75.1%; CG: 82%), while episode 9 had a mean comprehension of 91.5% (UG: 88.9%; CG: 93.1%). Results from a Wilcoxon signed-rank test revealed that comprehension was significantly higher in the last episode compared to the first episode viewed ($Z = 6.677$, $r = .61$, $p < .001$), and that this difference was significant both in the uncaptioned ($Z = 4.106$, $r = .61$, $p < .001$) and the captioned group ($Z = 5.316$, $r = .61$, $p < .001$).

Discussion

Comprehension of captioned and uncaptioned TV

Our first and second research questions investigated the extent to which adults understand captioned and uncaptioned episodes of the same television program over one semester of extensive viewing. Results from this study showed that participants had a mean comprehension of 73.5% across the nine uncaptioned L2 television episodes. Results also showed that comprehension varied significantly from episode to episode, oscillating between 65% to 89%. This indicated that even when comprehension was measured for the same group of participants and using the same type of test there were significant differences in comprehension amongst episodes of the same TV program. When captions were added, participants' mean comprehension was 86%, with mean scores ranging from 79% to 93%. Although there was less variation among the comprehension scores for the CG, differences in comprehension across episodes were also significant. These results suggest that comprehension of television programs is likely to be dependent on whichever episode of a television show is examined in a study, and this occurs for both captioned and uncaptioned viewing. Less variation across comprehension scores for the CG – indicated by overall smaller (medium-sized) effects – also indicates that the presence of captions may mitigate differences in comprehension across episodes.

The variation in viewing comprehension across episodes for both the captioned and uncaptioned groups has important research implications. First, comprehension of an individual episode is likely not representative of other episodes; if students can understand one episode of a television program, it does not ensure that they will understand another episode of the same program to the same degree. Thus, studies that investigate comprehension of a single episode may reveal the degree to which that episode is

understood. However, the findings of that study are likely not generalizable to other episodes of the same television program, nor would they likely be generalizable to episodes of other television programs. Second, because the results showed that participants' comprehension scores fluctuated depending on the episode, rather than improving progressively, researchers, teachers, and learners should not assume that if L2 learners understand one episode of a television program that their comprehension of subsequent episodes will improve. This contrasts Rodgers and Webb's (2011) suggestion that comprehension of L2 television programs is likely to improve through narrow viewing. However, it is possible that viewing nine episodes is not sufficient to reveal significant improvement in comprehension, and that more episodes are needed to capture development in comprehension through extensive viewing.

The mean percentages of comprehension for both uncaptioned (that is, 73.5%) and captioned (that is, 86%) conditions are similar to the findings of several earlier cross-sectional studies, which ranged from 40% to 75% for uncaptioned viewing (for example, Markham & Peter, 2003; Lee et al., 2021), and from 50% to 85% for captioned viewing (for example, Li, 2014; Matielo et al., 2017; Markham & Peter, 2003), although results fall in the high end of these spectrums. Higher mean percentage scores may be due to differences in participants' proficiency levels. For example, the participants in the present study were high-intermediate to advanced learners, whereas, the proficiency level of participants in the aforementioned studies were described as intermediate. However, if the comparison is made between the mean comprehension of any of the individual episodes from the present study and comprehension of a L2 television episode in a one-off study, results depend on which episode is selected. For example, Montero-Perez et al. (2014) found that participants' comprehension of uncaptioned television program clips was 68.1%, which is similar to the 67.3% achieved by participants in Episode 7. In contrast, in Matielo et al.'s (2017) study participants' mean comprehension of 20-minute episode was 76.6%, which is similar to the comprehension of Episode 1 (75.1%) in the present study. This constitutes further evidence that comprehension is highly dependent on the choice of audio-visual input, and that generalizations should be made with care.

To what extent do captions improve comprehension?

Our third research question explored the extent to which captions improve comprehension of episodes of the L2 television program. Results showed that participants who had access to captions consistently and significantly outperformed those learners viewing the episodes without the support of on-screen text. Participants in the captioned condition scored on average about 13% higher than participants in the uncaptioned condition, although differences ranged from around 5% to 19% depending on the episode. This indicates that the difference in comprehension is likely to vary across episodes of captioned and uncaptioned episodes of a television program, and suggests that studies that compare captioned and uncaptioned viewing of a single episode of a television program are unlikely to reveal the extent to which captions support comprehension. It might be assumed that as L2 learners view more episodes of a television program, the value of captions might diminish. However, the results also revealed that the support to comprehension provided by captions varied throughout the treatment, and it was not until the last episode that the added value of captions was shown to be non-significant.

Data revealed that the difference between conditions was significant in 8 out of the 9 episodes, and that it was non-significant only in the last episode (that is, episode 9). This result falls partly in line with two findings from Rodgers and Webb's (2017) study, who found (1) that comprehension fluctuated from episode to episode, regardless of the viewing condition, and (2) that differences were no longer significant by the 10th episode. This might indicate that through viewing 10 episodes of a single TV series, viewers may accumulate sufficient background knowledge and that makes the support provided by captions less necessary. In contrast with the results of the present study, however, Rodgers and Webb found that differences between the captioned and uncaptioned condition were – overall – non-significant. Two cross-sectional studies investigating viewing comprehension have also provided contrasting results. Lee et al. (2021) reported a non-significant advantage of viewing without captions over viewing with captions, while Lavaur and Bairstow (2011) found that viewing without captions led to significantly greater

comprehension than viewing with captions when participants had an advanced L2 level. However, the advantage of viewing with captions is supported by many earlier cross-sectional studies revealing that learners had superior L2 comprehension when viewing with captions than without captions, independently of differences in participants' L2 skills, background and study designs (for example, Birulés-Muntané & Soto-Faraco, 2016; Markham & Peter, 2003; Montero-Perez et al., 2014). Contrast in findings among studies indicates that further research examining comprehension of TV programs is needed, particularly using longitudinal designs that involve viewing large amounts of input over time.

Effects of vocabulary knowledge, lexical demands, and time

The fourth research question examined the extent to which learners' vocabulary knowledge, episodes' lexical demands, and time could predict comprehension. Results from a GLMM revealed that both vocabulary knowledge and lexical demands predicted comprehension rates, while time did not.

Vocabulary knowledge emerged as the second best predictor of comprehension after viewing condition. Independently of whether participants had access to captions or not, participants with greater vocabulary knowledge had higher comprehension scores. This should be expected, because greater vocabulary knowledge should allow viewers to understand a greater proportion of words encountered in L2 input. This in turn, should allow participants to devote more attentional resources to information processing (Webb & Nation, 2017). The lack of an interaction effect between vocabulary knowledge and viewing condition suggests that the presence of L2 captions does not affect the importance of learners' L2 vocabulary knowledge for viewing comprehension.

Unexpectedly, the level of lexical coverage provided by the first 1,000 word-families (plus marginal words) was negatively correlated with comprehension rates. This contradicts prior findings that have found that this variable has a small to medium sized correlation with viewing comprehension (for example, Durbahn et al., 2020; Pujadas & Muñoz, 2020; Rodgers, 2013). While an episode with a higher percentage of words belonging to the 1k word-family should theoretically have a lower lexical burden and be easier to understand than an episode with lower coverage, results from the present study do not support this assumption (nor did analysis of lexical coverage of the most frequent 2,000 and 3,000 word families). There may be several reasons for this. First, lexical coverage is one of many factors that affect comprehension and it may be that factors specific to spoken input (for example, speech rate, connected speech, clarity of speech), audiovisual input (familiarity with L2 audiovisual input, degree of congruency between imagery and words, amount of information provided by the imagery), and the materials (background information) had a large effect on comprehension. Second, measures of vocabulary knowledge may not adequately indicate the lexical coverage of input. Research investigating the effects of lexical coverage tend to use precise evaluations of learner knowledge of words encountered in the input by evaluating knowledge of all words in the input (Schmitt et al., 2011) or including pseudo-words in the input to gauge the proportion of unknown words (for example, Hu & Nation, 2000). However, there is no research explicitly examining the degree to which vocabulary test scores matched with a lexical coverage level predict comprehension (Webb, 2021). Thus, while the test scores in the present study matched against the lexical profiles of the episodes may provide an indication of lexical coverage, it may not be particularly accurate. Third, attaining the vocabulary knowledge necessary to match the lexical demands of an episode may indicate that a learner is more *likely* to understand it, but it does not ensure comprehension. Fourth, it may be that the participants were at a level where they all had sufficient lexical coverage of the episodes so its effects were marginal. Similarly, it may be that there was insufficient variation in the lexical coverage across the episodes to reveal its effects on comprehension.

Finally, the variable *time* (that is, the chronological order of the episodes) did not predict comprehension. Rodgers and Webb (2011) suggested that in a narrow viewing condition such as those used in this study, viewers may accumulate background knowledge and exposure to recurrent vocabulary as participants viewed more episodes of the same TV series and this may potentially ease comprehension. A direct comparison between the first and last episode showed that participants performed better in the last episode compared to the first, which falls in line with result from Rodgers and Webb's (2017) and

Fievez's (2020) studies. However, data also showed that there was no linear increase from episode 1 to 9. A lack of improvement over time was also reported by Pujadas and Muñoz (2020) and Gesa (2019), who did not find a linear improvement in comprehension in their 24-episode longitudinal studies. Pujadas and Muñoz suggested that comprehension might be episode-dependant, and the results of the present study provide support for this conclusion. However, it is also possible that more episodes might need to be viewed in order to see a clear improvement in comprehension over time, and that extensive/narrow viewing effects might not emerge through viewing just 9 episodes. It is also possible that the storyline of some episodes might have been intrinsically more difficult than others, interfering with this potentially 'linear' progression.

Conclusion

Findings from this study indicate that adult learners with high-intermediate to advanced L2 proficiency level could satisfactorily understand captioned and uncaptioned episodes from the same TV program. Comprehension varied significantly depending on the episode in both viewing conditions, suggesting that the fact that learners understand an episode does not mean they will understand another episode to the same degree. This has useful pedagogical implications, because it suggests that comprehension of a TV series cannot be assumed across episodes. While having access to captions seemed to level out differences amongst episodes, the extent to which captioned episodes were understood still varied significantly.

Regarding the effects of time, an improvement was observed from the first to the last episode. However, there was no linear progression. This suggests that L2 learners may need to view more than 9 episodes of television before seeing more transparent gains in comprehension. Results highlight the importance of longitudinal over cross-sectional studies to assess the long-term benefits of an extensive viewing approach. Studies looking into comprehension of different L2 television programs (for the same learners) would also offer valuable insights on how comprehension may vary across different types of television programs.

Finally, results of the present study suggest that learners' L2 vocabulary knowledge is a better predictor of comprehension than the episode's lexical demands. While lexical profiling can be useful to select materials of an appropriate level, larger or smaller vocabulary load may not accurately indicate the extent to which audio-visual materials will be understood. Research looking into vocabulary size and levels test scores matched against lexical profiles may help to clarify the degree to which different forms of assessment of lexical coverage predict comprehension (Webb, 2021).

Findings of this study should be considered in light of several limitations. First, it was a single-site study with one set of materials and tests, and it targeted a population that shared key demographic variables (for example, age, linguistic background). This limits the external validity of the findings, and generalizations to other individuals and contexts should be made with care (for example, American Psychological Association, 2018; Moranski & Ziegler, 2021). Nevertheless, the homogeneity of the sample (while including a wide range of L2-knowledge levels) mitigated potential background-related confounding effects (for example, participants had received similar L2-exposure in and outside the formal context). Considering the great diversity in TV programs and the language included in those programs, there is also likely to be variation in the extent to which viewing TV programs contributes to comprehension. Finally, the study did not consider variables such as episode topic complexity, the role of imagery, or learners' attention and enjoyment, which might have influenced comprehension. Future research looking into these factors may help explain the variability in comprehension across episodes.

Acknowledgements

We would like to thank the students and the professors that participated in the study. We also wish to extend our thanks to Emi Iwaizumi for her valuable advice on statistics. This study was funded by the Margarita Salas Grant for the training of young doctors from the Spanish Ministry of Universities.

Notes

1. Proper nouns represented, on average, an additional 3.64% of running words. Specifically, proper nouns added the following coverage to each episode: (Ep.1) +3.36%, (Ep.2) +2.20%, (Ep.3) +4.61%, (Ep.4) +4.48%, (Ep.5) +1.45%, (Ep.6) +3.42%, (Ep.7) +3.18%, (Ep.8) +4.33%, and (Ep.9) +5.56%.
2. The discrimination index was interpreted following the criteria suggested by Ebel and Frisbie (1972): very good (>.40), good (0,39 - 0,30), regular (0,29 - 0,20) and poor (<0,19).
3. A series of Shapiro-Wilk tests showed data was not normally distributed in any of the episodes, and five out of nine episodes presented unequal variances (see supplementary file).
4. Levene's test showed that variances for mean comprehension were unequal ($F(1,105) = 7.898$, $p = .006$), and data presented a skewed distribution.
5. The effect sizes for correlation coefficients were interpreted following the criteria suggested by Plonsky and Oswald (2014): small (.25), medium (.40), and large (.60).
6. See Note 4.
7. To further explore the effect of the episode's lexical demands, GLMMs were also run with other permutations of this variable (that is, lexical demands at 2,000-level, 3,000-level, with and without proper nouns), obtaining similar results.

References

- American Psychological Association (2018). External validity. In *APA Dictionary of Psychology*. <https://dictionary.apa.org/external-validity>
- Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching subtitled films can help learning foreign languages. *PLoS One*, *11*(6), 1–10. <https://doi.org/10.1371/journal.pone.0158409>
- Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International journal of listening*, *14*(1), 14–31. <https://doi.org/10.1080/10904018.2000.10499033>
- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 1–20. <https://doi.org/10.5334/joc.10>
- Buck, G. (2001). *Assessing listening*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732959>
- Charles, T. J., & Trenkic, D. (2015). Speech segmentation in a second language: The role of bimodal input. In Y. Gambier, A. Caimi, & C. Mariotti (Eds.), *Subtitles and language learning: Principles, strategies and practical experiences* (pp. 173–198). Peter Lang.
- Chung, J. M. (1999). The effects of using video texts supported with advance organizers and captions on Chinese college students' listening comprehension: An empirical study. *Foreign Language Annals*, *32*(3), 295–308. <https://doi.org/10.1111/j.1944-9720.1999.tb01342.x>
- Dizon, G., & Thanyawatpokin, B. (2021). Language learning with Netflix: Exploring the effects of dual subtitles on vocabulary learning and listening comprehension. *Computer Assisted Language Learning Electronic Journal*, *22*(3), 52–65. <https://callej.org/index.php/journal/article/view/352>
- Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, *88*, Article 102166. <https://doi.org/10.1016/j.system.2019.102166>
- Ebel, R. L., & Frisbie, D. A. (1972). *Essentials of educational measurement*. Prentice-Hall.

- Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, 42(3), 499–523. <https://doi.org/10.1017/S0272263119000494>
- Fievez, I. (2020). *What you see is what you get? Use and effectiveness of multimodal input for second language learning* [unpublished doctoral dissertation]. KU Leuven.
- Gesa, F. (2019). *L1/L2 subtitled TV series and EFL learning: A study on vocabulary acquisition and content comprehension at different proficiency levels* [unpublished doctoral dissertation]. Universitat de Barcelona.
- Gesa, F., & Miralpeix, I. (2024). Extensive viewing and L2 vocabulary learning: Two studies in EFL classes with children and adolescents. *International Journal of Applied Linguistics*, 175(2), 187–220. <https://doi.org/10.1075/itl.22013.ges>
- Guillory, H. (1998). The effects of keyword captions to authentic French video on learner comprehension. *CALICO Journal*, 15(1–3), 89–108. <https://doi.org/10.1558/cj.v15i1-3.89-108>
- Hayati, A. M., & Jalilifar, A. (2009). The impact of note-taking strategies on listening comprehension of EFL learners. *English Language Teaching*, 2(1), 101–111. <https://eric.ed.gov/?id=EJ1082250>
- Hayati, A., & Mohmedi, F. (2011). The effect of films with and without subtitles on listening comprehension of EFL learners. *British Journal of Educational Technology*, 42(1), 181–192. <https://doi.org/10.1111/j.1467-8535.2009.01004.x>
- Hsieh, Y. (2020). Effects of video captioning on EFL vocabulary learning and listening comprehension. *Computer Assisted Language Learning*, 33(5–6), 567–589. <https://doi.org/10.1080/09588221.2019.1577898>
- Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13, 403–430. <https://nflrc.hawaii.edu/rfl/item/43>
- Huang, H., & Eskey, D. (1999). The effects of closed-captioned television on the listening comprehension of intermediate English as a second language (ESL) students. *Journal of Educational Technology Systems*, 28(1), 75–96. <https://doi.org/10.2190/RG06-LYWB-216Y-R27G>
- Kelley, T. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30, 17–24. <https://doi.org/10.1037/h0057123>
- Klassen, K. (2021). Second language readers' use of context to identify proper names. *The Reading Matrix: An International Online Journal*, 21(2), 85–105. <https://mail.readingmatrix.com/files/25-9129tws5.pdf>
- Lai, H., Wang, D., & Ou, X. (2021). The effects of different caption modes on Chinese English learners' content and vocabulary comprehension. *International Journal of Computer-Assisted Language Learning and Teaching*, 11(4), 54–68. <https://doi.org/10.4018/IJCALLT.2021100104>
- Latifi, M., Mobalegh, A., & Mohammadi, E. (2011). Movie subtitles and the improvement of listening comprehension ability: Does it help? *The Journal of Language Learning and Teaching*, 1(2), 18–29. <https://dergipark.org.tr/en/download/article-file/209072>
- Laufer, B. (1989). What Percentage of Text-Lexis is Essential for Comprehension? In C. Lauren & M. Nordman, (Eds.), *Special language: From human thinking to thinking machines* (pp. 316–323). Multilingual Matters.
- Lee, M., & Révész, A. (2018). Promoting grammatical development through textually enhanced captions: An eye-tracking study. *The Modern Language Journal*, 102(3), 557–577. <https://doi.org/10.1111/modl.12503>

- Lee, P. J., Liu, Y. T., & Tseng, W. T. (2021). One size fits all? In search of the desirable caption display for second language learners with different caption reliance in listening comprehension. *Language Teaching Research*, 25(3), 400–430. <https://doi.org/10.1177/1362168819856451>
- Li, C. (2014). An alternative to language learner dependence on L2 caption-reading input for comprehension of sitcoms in a multimedia learning environment. *Journal of Computer Assisted Learning*, 30(1), 17–29. <https://doi.org/10.1111/jcal.12019>
- Markham, P., & Peter, L. (2003). The influence of English language and Spanish language captions on foreign language listening/reading comprehension. *Journal of Educational Technology Systems*, 31(3), 331–341. <https://doi.org/10.2190/BHUH-420B-FE23-AL>
- Markham, P., Peter, L., & McCarthy, T. (2001). The effects of native language vs. target language captions on foreign language students' DVD video comprehension. *Foreign Language Annals*, 34(5), 439–445. <https://doi.org/10.1111/j.1944-9720.2001.tb02083.x>
- Matielo, R., de Oliveira, R., & Baretta, L. (2017). Intralingual subtitles, interlingual subtitles, and video comprehension: insights from an exploratory study. *Letrônica*, 10(2), 758–774. <https://doi.org/10.15448/1984-4301.2017.2.26370>
- Mayer, R. (2001). *Multimedia learning*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139164603>
- Meara, P. M., & Miralpeix, I. (2006). *Y_Lex: The Swansea advanced vocabulary levels test*. v2.05. Lognostics.
- Meara, P., & Milton, J. (2003). *X_Lex: The Swansea levels test*. Express Publishing.
- Miralpeix, I. (2012, March). X_Lex and Y_Lex: A validation study. Paper presented at the 22nd Lexical Studies Conference, Newtown, UK.
- Montero-Perez, M., Peters, E., & Desmet, P. (2013). Is less more? Effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension. *ReCALL*, 26(1), 21–43. <https://doi.org/10.1017/S0958344013000256>
- Montero-Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*, 18(1), 118–141. <https://doi.org/10.125/44357>
- Moranski, K., & Ziegler, N. (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning*, 71(1), 204–242. <https://doi.org/10.1111/lang.12434>
- Muñoz, C. & Miralpeix, I. (2024) *Audiovisual Input and Second Language Learning*. John Benjamins. <https://doi.org/10.1075/llt.61>
- Nation, P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [Software]. <https://www.victoria.ac.nz/lals/about/staff/paulnation>
- Nurmukhamedov, U., & Webb, S. (2019). Lexical coverage and profiling. *Language Teaching*, 52(2), 188–200. <https://doi.org/10.1017/S0261444819000028>
- Oppenheimer, J., & Arnaz, D. (1951) *I love Lucy* [TV series], CBS.
- Pattemore, A., & Muñoz, C. (2020). Learning L2 constructions from captioned audio-visual exposure: The effect of learner-related factors. *System*, 93, Article 102303. <https://doi.org/10.1016/j.system.2020.102303>

- Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly*, 53(4), 1008–1032. <https://doi.org/10.1002/tesq.531>
- Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, 40(3), 551–577. <https://doi.org/10.1017/S0272263117000407>
- Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: A study of L2 vocabulary learning by adolescents. *The Language Learning Journal*, 47(4), 479–496. <https://doi.org/10.1080/09571736.2019.1616806>
- Pujadas, G., & Muñoz, C. (2020). Examining adolescent EFL learners' TV viewing comprehension through captions and subtitles. *Studies in Second Language Acquisition*, 42(3), 551–575. <https://doi.org/10.1017/S0272263120000042>
- Pujadas, G., & Muñoz, C. (2023). Measuring the visual in audio-visual input: The effects of imagery in vocabulary learning through TV viewing. *International Journal of Applied Linguistics*, 174(2), 263–290. <https://doi.org/10.1075/itl.22019.puj>
- Pujadas, G. & Muñoz, C. (2024). When to switch captions off? Exploring the effects of L2 proficiency and vocabulary knowledge on comprehension of captioned and uncaptioned TV. *Studies in Second Language Learning and Teaching*, 14(3), 545–570. <https://doi.org/10.14746/ssllt.38036>
- Rodgers, M. (2013). *English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes, and captions* [Unpublished doctoral dissertation]. Victoria University of Wellington.
- Rodgers, M. P., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, 45(4), 689–717. <https://doi.org/10.5054/tq.2011.268062>
- Rodgers, M. P., & Webb, S. (2017). The effects of captions on EFL learners' comprehension of English-language television programs. *CALICO Journal*, 34(1), 20–38. <https://www.jstor.org/stable/90014676>
- Rodgers, M. P., & Webb, S. (2020). Incidental vocabulary learning through viewing television. *International Journal of Applied Linguistics*, 171(2), 191–220. <https://doi.org/10.1075/itl.18034.rod>
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26–43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, 41(3), 609–624. <https://doi.org/10.1016/j.system.2013.07.012>
- Vulchanova, M., Aurstad, L. M., Kvitnes, I. E., & Eshuis, H. (2015). As naturalistic as it gets: subtitles in the English classroom in Norway. *Frontiers in Psychology*, 5, 1–10. <https://doi.org/10.3389/fpsyg.2014.01510>
- Wang, A., & Pellicer-Sánchez, A. (2022). Examining the effectiveness of bilingual subtitles for comprehension: An eye-tracking study. *Studies in Second Language Acquisition*, 45(4), 882–905. <https://doi.org/10.1017/S0272263122000493>
- Webb, S. (2011). Selecting television programs for language learning: Investigating television programs from the same genre. *International Journal of English Studies*, 11(1), 117–135. <https://doi.org/10.6018/ijes/2011/1/137131>
- Webb, S. (2015). Extensive viewing: Language learning through watching television. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 159–168). Routledge.

- Webb, S. (2021). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, 33(2), 287–302. <http://hdl.handle.net/10125/67407>
- Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.
- Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, 14(1), 65–86. <https://doi.org/10125/44203>
- Wisniewska, N., & Mora, J. C. (2020). Can captioned video benefit second language pronunciation? *Studies in Second Language Acquisition*, 42(3), 599–624. <https://doi.org/10.1017/S0272263120000029>

Appendix A. Studies Comparing Comprehension of Captioned and Uncaptioned Videos

Study	Learners' information				Study design			Summary of results			
	L1 ^a	Age	TL ^b	TL proficiency (test type, if specified)	N	Video material	Testing format ^c	Captions	No Captions	Diff. ^d	Effect size ^e
Birulés-Muntaner & Soto-Faraco (2016)	Cat Spa	21-28	En	B2 (placement test)	60	60-min full episode	8 items <i>format n/a</i>	79.4%	59.4%	20%**	$d = 1.37$
Chung (1999)	Chi	17-19	En	low-intermediate	183	video fragment (<i>length n/a</i>)	10 items MC (4 opt, L1)	76.6%	66.9%	9.7%**	$g = 0.682$
Guillory (1998)	En	Uni ^f	Fr	N/A	202	two video fragments (<i>length n/a</i>)	2 tests x 7 items short answer (L1)	71.8%	52%	19.8%**	N/A
Hayati & Mohmedi (2011)	Per	$M: 22$	En	Intermediate (English Language Proficiency Test)	90	six 5-min fragments	6 tests x 10 items MC (L1)	66.3%	41%	25.3%**	$g = 5.375$
Huang & Eskey (1999)	N/A	N/A	En	Intermediate (placement test)	30	14-min episode (viewed twice)	16 items MC (3 opt, L1)	67.9%	47.9%	20%**	$g = 1.293$
Hsieh (2020)	Chi	Uni	En	low-intermediate (TOELIC/TOEFL)	105	two 4-5-min animation clip	2 test x 10 items MC	65.7%	60.25%	5.45%	N/A
Latifi, et al. (2011)	Per	17-30	En	Intermediate (IELTS)	36	fifteen 2-min fragments	15 tests x 10 items MC	66.3%	52.5%	13.8%*	$\eta^2 = 0.218$ ^g
Lavaur & Bairstow (2011)	Fr	15-18	En	Beginner (B) Intermediate (I) Advanced (A)	90	8-9-min clip	42 items	B 32.8% I 55.7% A 74.8%	22.8% 54.8% 85.7%	10% 0.9% -10.9%* ^h	$\eta^2 = 0.432$ ^g
Lee, et al. (2021)	Chi	Uni	En	B2 (TOELIC/TOEFL)	95	11-min TED talk	15 items MC	74.2%	74.9%	-0.7%	N/A
Li (2014)	Chi	Uni	En	Intermediate (TOELIC)	97	20-min episode	18 items MC (4 opt, L1)	84.7%	37.9%	46.8%**	$\eta^2 = 0.85$ ^g
Markham, et al. (2001)	En	Uni	Spa	Intermediate	169	7-min episode	10 items MC (L1)	56.7%	46.1%	10.6%**	$g = 0.548$
Markham & Peter (2003)	En	Uni	Spa	Intermediate (‘last course grade’)	213	7-min episode	20 items MC (4 opt, L2)	50.6%	39%	11.6%**	$g = 0.967$

Matielo, et al. (2017)	Pt	<i>M</i> : 22	En	Intermediate	36	20-min episode	8 items ⁱ OE, TF	G ⁱ 75.0% S ⁱ 86.6%	53.3% 76.6%	21.7% 10%	$\omega = 0.19$ ^j
Montero-Perez, et al. (2013)	Dut	<i>M</i> : 19	Fr	Intermediate (VLT ^k)	226	3 clips (8, 6, 2 min. respectively)	43 items mix formats (L1)	59.6%	54.2%	5.4%**	$\eta^2 = 0.05$ ^g
Montero-Perez, et al. (2014)	Dut	<i>M</i> : 18	Fr	high-intermediate (VLT)	133	3 clips (2, 4, 3 min. respectively)	41 items OE, MC and COMB (L1)	71.8%	68.1%	3.7%	$\eta^2 = 0.01$ ^g
Pujadas & Muñoz (2024)	Cat Spa	Uni	En	A1 – C2 (OPT test)	250	nine 30-min episodes	9 tests, 90 items MC (3 opt.), TF (L1)	87.4%	73.7%	13.7%**	$r = 1.39$
Rodgers (2013) Rodgers & Webb (2017)	Jpn	Uni	En	pre-intermediate to intermediate (VLT)	488	ten 43-min episodes	10 tests, 742 items MC (3 opt.), TF and SEQ (L1)	66.5%	63.9%	2.6% ^l	η^2 (e1) = 0.046 η^2 (e4) = 0.031 η^2 (e7) = 0.013
Wang & Pellicer-Sánchez (2022)	Chi	18-34 <i>M</i> : 23	En	B2 – C1 (IELTS)	112	Four clips (23 min. total)	34 items MC (4 opt., L1)	69.1%	60.2%	8.9%	$d = 0.69$

^a L1 = First Language: Cat (Catalan), Chi (Chinese), Dut (Dutch/Flemish), En (English), Fr (French), Jpn (Japanese), Per (Persian/Farsi), Pt (Portuguese), Spa (Spanish)

^b TL = Target Language: En (English), Fr (French), Spa (Spanish)

^c MC = Multiple-choice; OE = Open-ended; COMB = combination (e.g., match sentence and picture); TF = True-False; SEQ = sequencing (e.g., order events chronologically). When specified, the language in which the items were presented has been included (i.e. L1, L2)

^d Diff. = Difference between the Caption group and the No Caption group. To facilitate comparison, the difference is displayed in percentage – if not available, this was been calculated based on the mean score provided and the number of items. Note that only results from the captions and non-captions group are reported, although some of the studies included other experimental conditions such as key-word captioning, L1 subtitles – which also outperform the no-text conditions – or advanced organizers.

^e Effect sizes are presented in the reported standardized measure when available (i.e. Cohen's *d*, eta-squared (η^2), partial eta-squared (η^2_p)). Hedges' *g* has been reported for those studies included in Montero et al.'s (2013) meta-analysis, and η^2 has been calculated for those studies that had not reported the effect size for ANOVAs but provided enough information to obtain it.

^f Uni = University students

^g The study included other experimental groups (e.g. L1-subtitles, key-word captioning), and thus the effect size reported corresponds to the variance explained by experimental condition, not by the addition of captions. Nevertheless, the study reports that captioned viewing outperformed uncaptioned viewing, independently of the performance of other experimental groups.

^h Data corresponds to questions based on dialogue information only. The study also included items based on visual information, which are not discussed here.

ⁱ Two type of questions were used: General (G) and Specific (S).

^j Authors calculated the effect size by taking the chi-square value divided by $n-1$.

^k Vocabulary Levels Test

^l Rodgers (2013) and Rodgers & Webb (2017) found that the mean difference from the 10 episodes was not significant. It reached statistical significance, however, in three of them (with a mean difference of 4.5% in those three). The effect sizes reported correspond to those 3 episodes (e1 = Initial Episode; e4 = Episode 4; e7 = Episode 7)

* significant at the .05 level.

** significant at the .001 level.

Appendix B. Examples of items' type and format

Textually explicit items (in MC and TF format)

What does Lucy want to get from John Wayne's dressing room?

- a. A poster with his signature.
- b. The shoes he wore in Blood Alley.
- c. **A pair of boots.**

SCRIPT: 'Ethel, we're saved. The signature, it isn't touched, see? All I have to do now is go in there, **steal a pair of his boots**, make the imprints and our troubles are over'.

V / **F** John Wayne's friends call him 'Count'.

SCRIPT: 'What's the matter with me? Why don't I call Duke? Who? Duke - John Wayne. **All his friends call him Duke.**'

Inferential items (in MC and TF format)

Why didn't Ricky buy the Mertzes' tickets?

- a. Because he knew the Mertzes wanted to travel with their motorbike.
- b. Because he thought the Mertzes would buy their own tickets
- c. **Because he had a lot of things in his mind.**

SCRIPT: 'I bought the train tickets. How many? Three. Only three? Yeah. The baby doesn't need one. One for your mother, one for you and one for me. Oh, no. The Mertzes. Oh, yes, the Mertzes. What's the matter with you, Ricky? Well, I don't know. I was busy thinking of that travel family plan and try to save some money and the car.'

V / **F** Fred has gone fishing three times without Ethel.

SCRIPT: 'Fred, in all the years you've been married, have you ever left Ethel alone to go fishing? Nope. You see? He loves his wife.'

Note. Comprehension questions / answers have been translated to English for comprehensibility.

Appendix C. Generalized Linear Mixed Model

Model Summary

Model	AIC	BIC	Deviance	df	X ²	<i>p</i>
H_0 (random effects only)	7946	7967	7940			
H_1 (full model)	7855	7898	7843	3	96.3	< .001

Results of Generalized Linear Mixed-Effects Modeling

Fixed effects	Estimate	SE	<i>z</i>	Odds Ratio	95% CI		<i>p</i>
					Lower	Upper	
Intercept	1.387	0.139	10.00	4.00	3.05	5.26	< .001
Captions (yes)	0.875	0.103	8.52	2.40	1.96	2.94	< .001
Vocabulary knowledge	0.333	0.052	6.34	1.39	1.26	1.55	< .001
Lexical coverage	-0.427	0.119	-3.58	0.65	0.52	0.82	< .001

Random effects	Variance	<i>SD</i>	ICC
Participants	0.185	0.43	0.040
Item	1.153	1.07	0.249

Conditional $R^2 = .357$; Marginal $R^2 = .095$

Estimated Probabilities (categorical factor)

Group	Probability	SE	95% CI	
			Lower	Upper
Captions	0.91	0.011	0.88	0.93
No captions	0.80	0.022	0.76	0.84

About the Authors

Geòrgia Pujadas is a postdoctoral researcher at the University of Western Ontario and the University of Barcelona. Her research interests include EFL learning through audiovisual input, vocabulary acquisition, and listening comprehension. Her research has been published in journals such as *Studies in Second Language Acquisition* and *Second Language Research*.

E-mail: georgia.pujadas@ub.edu

ORCID: <https://orcid.org/0000-0002-0290-1158>

Stuart Webb is a Professor of Applied Linguistics at the University of Western Ontario. He currently teaches on the Masters in TESOL program and supervises students at the MA and PhD levels. His latest books are *The Routledge Handbook of Vocabulary Studies*, and *How Vocabulary is Learned* (with Paul Nation).

E-mail: swebb27@uwo.ca

ORCID: <https://orcid.org/0000-0002-8297-4997>